SMU Data Science Review

Volume 3 | Number 2 Article 14

2020

Automated Interactive 3D Geospatial Data Assimilation, Formatting and Visualization System for Development of Subsurface Conceptual Site Models

AARON CATTLEY
Southern Methodist University, acattley@mail.smu.edu

Gavin Hudgeons
ghudgeons@mail.smu.edu

Bruce Lee
brucel@mail.smu.edu

Automated Interactive 3D Geospatial Data Assimilation, Formatting and Visualization System for Development of Subsurface Conceptual Site Models

Gavin Hudgeons, Aaron Cattley, Bruce Lee, and John Carney

Master of Science in Data Science, Southern Methodist University, Dallas TX 75275 USA

Abstract. The natural evolution of the collection and storage of subsurface data in Texas has resulted in the current state where data for certain resources, such as water resources, have not been assimilated with state oil and gas and injection data in a meaningful way that allows for rapid understanding and data analysis for a physical land site. The consequences that result due to data from different spheres not being in sync are often duplication of work being performed but not in a consistent manner. However, the reality is that the infrastructure and impacts of these sectors are deeply intertwined. Lack of understanding of each sector can lead to massive revenue loss and damage to natural resources. Even the regulatory agencies who issue the permits designed to protect resources are not aware of what is going on. Additional consequences that result from not being on the same page are duplication of unnecessary expensive permits, failed groundwater protection and freshwater exploitation. What we propose to develop is a midstream product that assimilates these disparate data resources into an end-user centric data assimilation, visualization, and analysis service. This product would not only provide a one stop solution for customers, but also alleviate wasted costs and help preserve the environment.

1 Introduction

The ability to demonstrate comprehension of complex subsurface geospatial data is a fundamental element of successful earth science consulting and engineering. Seeing multiple data sets in their true geospatial context is a powerful way to understand and communicate the intricacies of a site. Indeed, if a conceptual site model (CSM) is the synthesis of assimilated data into a graphical representation, a 3D or 4D visualization can itself become the conceptual model. A 3D visualization is a representation of site data in a realistic or relative context to its actual geospatial location. A 4D visualization includes the fourth dimension of time and is therefore a transient depiction of a 3D model or, more simply, a 3D animation.

3D visualizations are not new. Geologists have illustrated their data with 3D drawings and models for centuries. What have changed are the tools. What was

once pen and ink are now geospatial visualization technologies that allow us to rapidly create hundreds of multidimensional views of a site, animated over multiple parameters and digitally distributable in an interactive format. Data scientists now work with innovative technologies specifically designed to capture and assimilate vast amounts of geospatial site data and render those data into interactive 3D or 4D visualizations. An expanding assortment of visualization software is available to address the needs of the working data scientist. CTech's Earth Volumetric Studio®, Esri® ArcGIS® 3D Analyst®, RockWare® RockWorks®, and EarthVision® by Dynamic Graphics, Inc., are but a few examples of some commonly used commercial geospatial visualization applications.

Because all geospatial data have the same basic structure (X, Y, Z, n1, n2, n3,...), visualization techniques are highly adaptable from one application to another. For example, output from multiple geostatistical modeling packages can be exported into a single common visualization engine for rendering, further analysis, or delivery. The same technologies used by doctors to visualize brains can be used by geologists to visualize aquifers. The ability to move between packages taps the open-source nature of geospatial data, expanding the toolbox and simplifying collaboration.

3D visualization is not limited to static 3D views. Instead, multidimensional, high-resolution, multi-frame, interactive, comprehensive, portable, and distributable visualizations are able to be developed that assimilate massive amounts of site data into intuitive visual systems. They are used to demonstrate both concrete and abstract ideas and represent legacy, real-time, and forward modeling data. 3D conceptual site model visualizations (CSMVs) are used for the following purposes:

- Communication
- 2D and 3D paper graphics production
- Training and educating
- Project management
- Performance tracking and quality control/quality assurance (QA/QC)
- Data assimilation

CSMVs are built with flexible interfaces and adaptable workflows designed to allow for rapid and effective deployment of 3D or 4D to meet one's programmatic needs. Every geospatial CSM is unique. No two sites demand exactly the same sequence of visualizations to be built, interpolations to be performed, or animations to be rendered. However, because of the common structure of geospatial data, many commonalities exist between all forms of 3D visualization.

In addition, 3D data availability has exploded. With online government geospatial databases, web map services, and geospatial depots, more data are readily available to be visualized than ever before. The web has brought much of the world's subsurface data into the digital age. Groundwater well and oil well data sets are ever-expanding. Remote sensing and geophysical data collection techniques are increasing in both coverage and resolution. Land use, boundary and infrastructure vectors are readily available in useable geospatial formats.

Governments have invested significant resources in the development of online geospatial data repositories, and we are the beneficiaries.

However, with this explosion of online data availability, the natural evolution of the collection and storage of subsurface data in Texas has resulted in our current state where data for certain resources, such as water resources, have not been effectively assimilated with State oil, gas, and injection data in a meaningful way that allows for rapid understanding and data analysis for a site. Further, many of the public Texas databases, particularly those maintained by the Texas Railroad Commission, use antiquated systems that make it difficult for typical user to access and download large dataset queries in a usable format. Further still, each repository has its own standards on data formats, key identifiers, and coordinate system projections. The result of these deficiencies have resulted in knowledge gaps that can have detrimental consequences associated with natural resource management, such as the contamination of groundwater resulting from improper disposal of produced water generated from oil and gas wells [1].

Further, commercial companies who traditionally may have been thought as being the ones to assimilate and standardize data from these repositories have let many customers down, leaving a gaping hole in the market for a new product to fill. Two such companies include IHS Markit and Enverus. While both companies offer feature-rich products, they are often bloated and prohibitively expensive for the needs of a general consumer.

A solution to the conditions described above is the development of an automated geospatial data aggregation and visualization system. The system relies on free and publicly available database resources maintained by State and Federal government agencies and produces output that is useful to the end-user without the need for expensive software licenses. Thus, this paper describes the development of such a system which assimilates and visualizes these disparate data resources into an end-user centric 3D conceptual site model (3D CSM), hereafter referred to as 'the product'. One way to think about the product is as a 'Google Earth for the Subsurface'. Further, the product is also an analytical tool that is customized for the end user's specific site and is delivered, on-demand, with 3D visualizations and data formatted into GIS, machine learning, and end-user friendly formats.

2 Development of Geospatial Conceptual Site Models with Respect to the Product

The purpose of the product is to automate the generation of geospatial subsurface 3D CSMs. The input given to the product by the user is a simply a bounding box, consisting of a maximum longitude, a minimum longitude, a maximum latitude, and a minimum latitude. With this input the product is designed to 'scrape' online data repositories for the geospatial data that exists within the

[1] More information may be found at https://www.texastribune.org/2016/08/24/texas-promised-34-years-ago-track-oilfield-waste-a/. Last accessed 14 Aug. 2020.

defined bounding box necessary for the development of a 3D CSM. The product then formats the data into suitable formats for a 3D visualization system. The product takes advantage of CTECH Earth Volumetric Studio (EVS) for visualization. EVS is a powerful visualization system with a complete integrated earth science toolkit. EVS also has the advantage that its tools can be commanded with Python and has libraries to output interactive 3D visualizations that can be deployed without end users requiring a license purchase. As such, data obtained from the repositories are formatted into formats usable by EVS. A more detailed description of the data formats used for the product are discussed in Section 2.3.

What follows is a discussion of both a generalized approach to the development of 3D CSMs from geospatial subsurface data, and how specific elements from this approach were used for the development of the product.

A 3D CSM can be thought to exist within a digital 3D geospatial framework. The framework establishes the space within which assimilated data can be visualized in a uniform format. Optimally, the framework consists of a single site-wide verified data set visualized using software that relates the spatial data. For this reason, site-wide digital ground elevation data sets, or digital elevation models (DEMs) make good geospatial frameworks. Usually the framework data set is the first data set to be visualized. The concept of a framework data set is analogous to a basemap in 2D mapping applications. In general, the framework data set is designed to:

- Define the coordinate extents of the model
- Utilize both site-specific and large geographic areas as needed
- Prepare the space for the assimilation of all the appropriate site data into a comprehensive and uniform visual format
- Provide a reliable data set for comparative visual QA/QC of forthcoming data
- Easily assimilate new data as they become available

A simple visualization framework might consist of a 3D topographic surface defining the X and Y extents of a model. The coordinate limits defined for the framework dataset are selected specifically to accommodate future data sets to be incorporated. Usually this is defined as the area of interest (AOI).

For the purposes of the product, the framework dataset is a DEM that defines a topographic surface. A topographic surface is usually the highest-resolution layer of a subsurface 3D CSM. For large-scale regional models with high-resolution data, the framework may be designed to automatically adjust resolution with zoom levels to accommodate workstation limitations. Figure 1 shows an example of a tiled mosaic of National Elevation Dataset (NED) DEM visualized as a topographic surface as the basis for a regional visualization framework. Figure 1 also shows four NED tiles stitched together and visualized in 3D at full resolution. For each tile, a consistent color scheme has been applied that accounts for the minimum and maximum extents of elevation for the combined visualized tiles. Larger areas can be shown, but resolutions may need to be reduced to accommodate workstation limitations. Higher-resolution Light Detection and Ranging (LiDAR) topographic data, visualized in Figure 2, may

also be incorporated as available. The product relies on 1/3 arc-second DEM datasets from the United States Geological Survey (USGS). These DEMs were selected because they represent the highest resolution seamless DEM datasets available for the United States.

Aerial photos and maps can also be rasterized and georeferenced to be draped onto any related 3D surface and included in the database to provide reference and annotations to the surface. Figure 4 shows an example of five paper maps, an aerial photograph, and four tiles of NED data in 3D geospatial synergy. Frame-by-frame 4D animations chronologically scrolling through legacy paper documents draped onto a 3D topographic surface can be a quick and effective way to display historic site conditions and interpretations. The product takes advantage of the National Agriculture Imagery Program (NAIP) aerial photos maintained by the USGS.

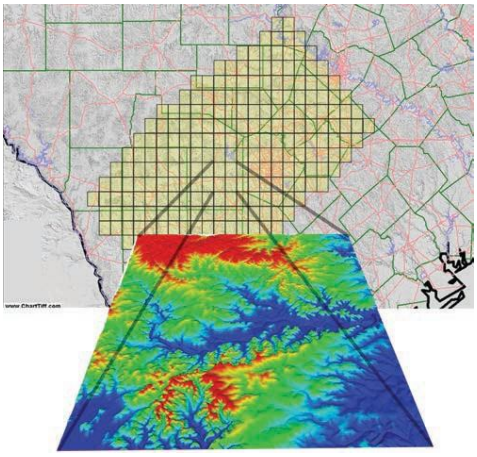


Fig. 1. Example of how maps may be 'draped' onto 3D topographic surfaces for annotation. This figure shows five paper maps, an aerial photograph, and four tiles of NED data in 3D geospatial synergy.

Once a 3D visualization framework is designed and annotated, it can be populated with project data. A relational visualization database (RVD) is designed for CSM development. Data incorporated into the RVD include all pertinent geospatial site data, both current and legacy, within the coordinate extents of the visualization framework.

The RVD may include digital and tabular data as well as scanned paper maps and images. Common data types include well logs, geologic descriptions, geophysical data, soil, 4D surface and groundwater chemistry data, 4D potentiometric surface data, oil and gas production and attribute data, preexisting maps, preexisting GIS and geospatial databases, cultural data, surface and infrastructure data, and conceptual data. Except in cases where on-the-fly coor-

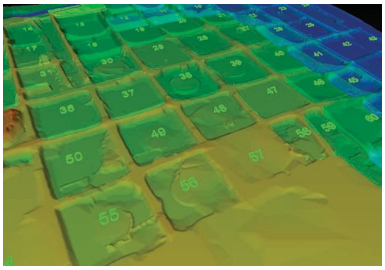


Fig. 2. Example of a 3D Visualization of higher resolution LiDAR data.

Fig. 3. Example of how maps may be 'draped' onto 3D topographic surfaces for annotation. This figure shows five paper maps, an aerial photograph, and four tiles of NED data in 3D geospatial synergy.

dinate projections will be used, all data in the RVD must be defined within a consistent coordinate system, however, it is advisable to always use one uniform coordinate system as many programs for spatial data analysis cannot project on the fly. At the time of this publication, the product currently takes advantage of the following datasets to populate CSM RVDs:

- Texas Water Development Board (TWDB) Groundwater Database
- United States Geological Survey provided NAIP Imagery and National Elevation Dataset (NED) Digital Elevation Models (DEMs)
- Texas Railroad Commission Underground Injection Control (UIC) Database
- Texas Railroad Commission Digital Map Datasets
- Texas Railroad Commission Oil and Gas Well Attribute Datasets

Once the data have been formatted, they are read into EVS where the visualizations are created, optimized, and rendered for delivery. EVS provides a bounty of visualization settings and geostatistical interpolation algorithms for the development of 3D CSMs. Many different techniques are used to visualize 3D CSMs. For example, elements of geologic data can be visualized, from

points to lines to surfaces to volumes. Traditional geologic mapping techniques, such as cross sections, fence diagrams, and isopleth maps, translate naturally into 3D. Stratigraphic layers can be stretched apart or turned on and off like lights. Faults can be visualized as complex 3D surfaces, and fault blocks can be displaced, uplifted, and eroded through 3D animations. Slice planes can be run through models in any direction. Multiple slice planes can be positioned anywhere, and 3D well-to-well fence diagrams can be cut from a model.

One can apply custom color scales, display contours, or make geologic layers pinch out and disappear below specified thicknesses. Vertical exaggeration can be applied to discern subtle topographic features or for projects that have a large horizontal span relative to depth. Transparency values, lighting effects, and visual effects, such as volume rendering, can be customized. Any parameter can be animated, from benzene concentrations over time to stratigraphic thickness isovolume levels. Figure 4 shows several views of a geologic model. Figure 4 is a kriged geologic model based on correlated boring data.

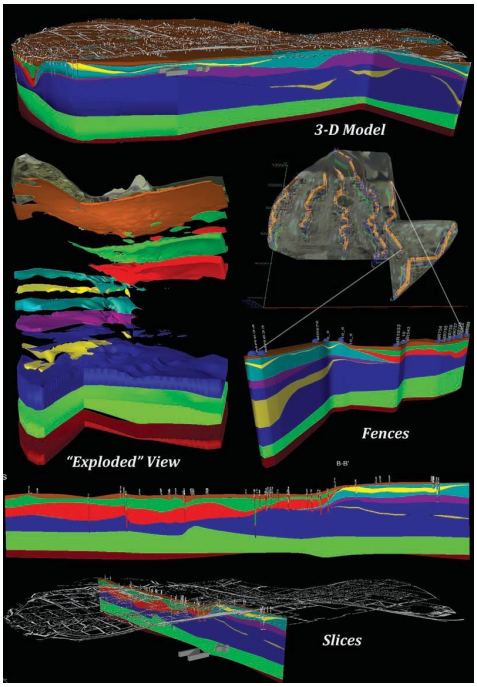


Fig. 4. Examples of 3D manipulation of a geologic model, including a full volumetric model, an 'exploded' view where layers of the model are separated, fence diagrams, and slices.

3D CSMs are also powerful tools for revealing trends in groundwater conditions. Figure 5 depicts frames captured from a visualization of a groundwater

remediation system. Incorporating thousands of groundwater samples collected over a decade, concentrations were visualized tracking plume remediation conditions relative to groundwater elevations. Well-construction details, sampling locations, property boundaries, and aerial photographs are also incorporated. A flight beneath the site reveals the effect of the pump-and-treat system on groundwater elevations and chemical concentrations through time while simultaneously displaying daily sampling plans (purple balls) with dynamic titles.

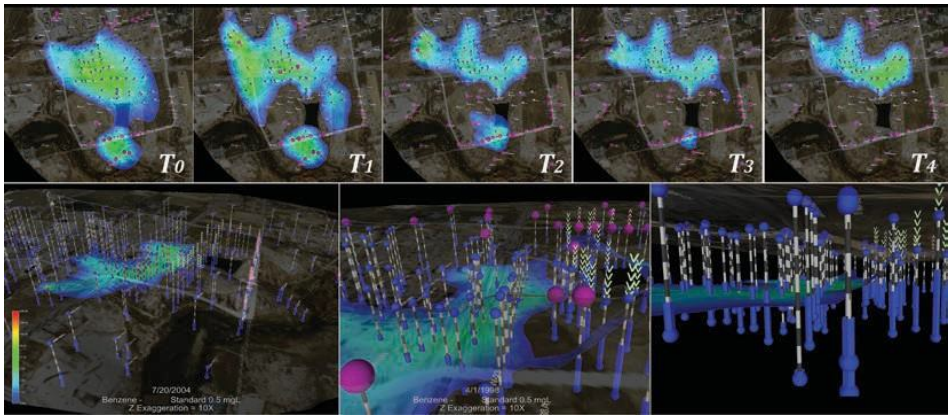


Fig. 5. 3D Visualization of Groundwater Chemistry. This figure depicts results of thousands of groundwater samples collected over time and visualized in a 3D geospatial domain with associated well borings, DEM surfaces, and surface imagery.

3D visualization of subsurface soil and groundwater data is useful for many purposes, including calculating the volume of contaminated groundwater and soil and the overall contaminant mass in the subsurface, performing spatial moment calculations, such as the center of mass and the plume spread, visualizing the spatial relationship of the plume to receptors, identifying preferential pathways, and verifying the horizontal and vertical nature and extent of contamination. 4D animations of the above parameters can help visualize changes over time and are very useful in monitored natural attenuation studies. Figure 6 shows a visualization of a 3D soil plume kriged from chemical samples collected during a boring investigation. Frames captured from two types of animations are shown: a concentration-based animation whereby each frame depicts plume concentrations at increasing isolevels and a contaminant-based animation whereby each frame represents a different chemical of concern (COC) at the appropriate regulatory levels. Higher-resolution geophysical log data, CPT, or rapid optical screen tool data can also be visualized (Figure 7). Because remote sensing and geophysical data are often collected in very fine vertical intervals (centimeter scale or less), vertical exaggeration must often be used to visualize subtle vertical changes, and

grids must be of sufficient resolution to capture these variations, which are often critical to the CSM.

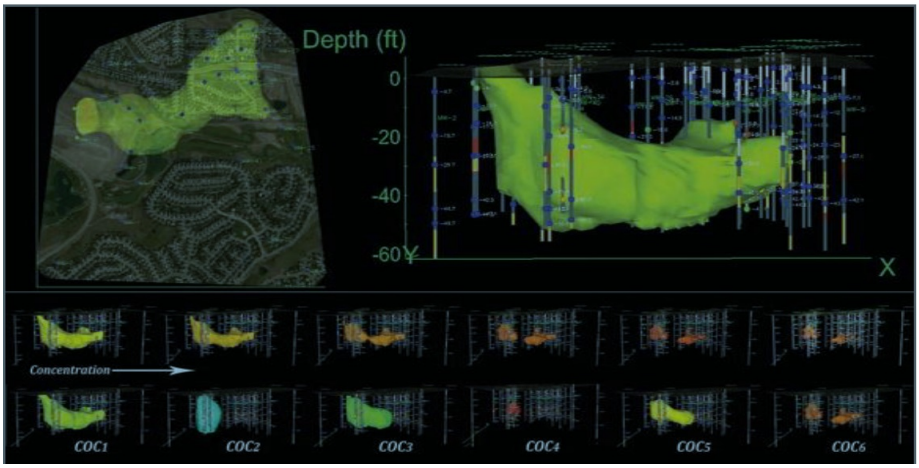


Fig. 6. 3D Visualization of a 3D soil plume kriged from chemical samples collected during a boring investigation. Frames captured from two types of animations are shown: a concentration-based animation whereby each frame depicts plume concentrations at increasing isolevels and a contaminant-based animation whereby each frame represents a different chemical of concern (COC) at the appropriate regulatory levels.

Because the product is an automated 3D CSM generator, decisions had to be made ahead of time as to which of these types of features to make available to the end user. Fortunately, the May 2020 release of EVS provides a delivery format, termed 'EVS Presentations', that put control of many of these features into the hands of the end user without the need for a license purchase. As such, the product is designed to give control of the following 3D CSM settings to the end user: vertical exaggeration, slice Plane Positioning, and attribute visibility.

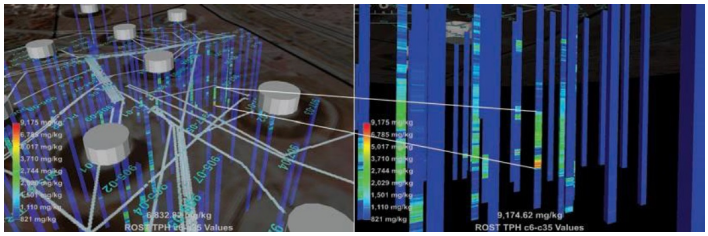


Fig. 7. 3D Visualization of high-resolution geophysical log data.

Although these settings are being put into the hands of the end-user, it is recognized that these settings must be used with care. Figure 8 shows an extreme use of vertical exaggeration. 2D road maps, shown in red, were draped onto LiDAR data over a large regional area. The LiDAR data are turned off, leaving the 2D roads with their newly assigned third dimension. A duplicate of the roads created 2 ft down, shown in black, is the compaction zone. The spatial relationship between the compaction zone and the regional potentiometric surface, in blue, is shown. The vertical exaggeration is 200 times, which allows us to visually perform this 2-ft vertical analysis beneath a 1.3-mi2 region in a single view. While this represents an appropriate use of vertical exaggeration, for most applications, this extent of exaggeration would be misleading and result in erroneous data interpretations. In some cases, excessive exaggeration may be used for manipulative purposes, similar to the intentional modification of the axes of graphs to mislead the reader. In general, the end user must be mindful of vertical scale when applying settings to the 3D CSMs and maintain ethical standards in his or her own practice.

Once the data have been assimilated in the RVD, a set of 3D visualizations can be developed to visually apply Quality Assurance/Quality Control (QA/QC) to the data for any incorrect or misplaced elements. Figure 9 shows boring coordinate errors detected by visualization QA/QC.

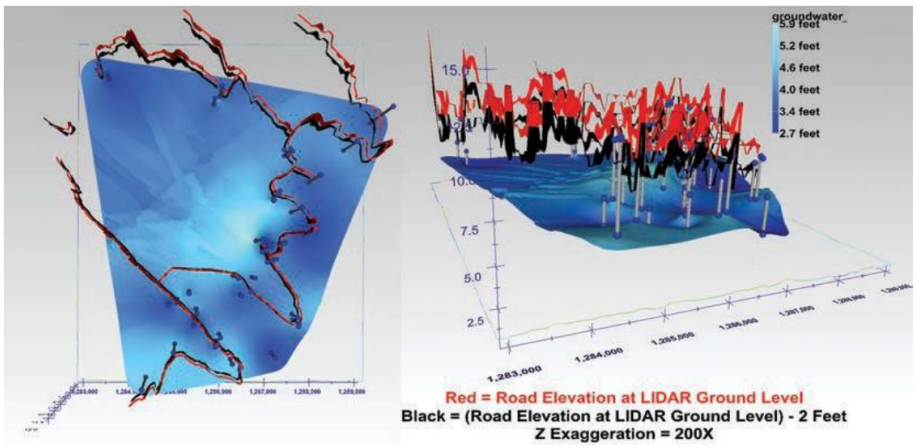


Fig. 8. Example of an extreme use of vertical exaggeration. For most applications, this extent of exaggeration would be misleading and result in erroneous data interpretations. End users must be mindful of vertical scale when applying settings to the 3D CSMs and maintain ethical standards in his or her own practice.

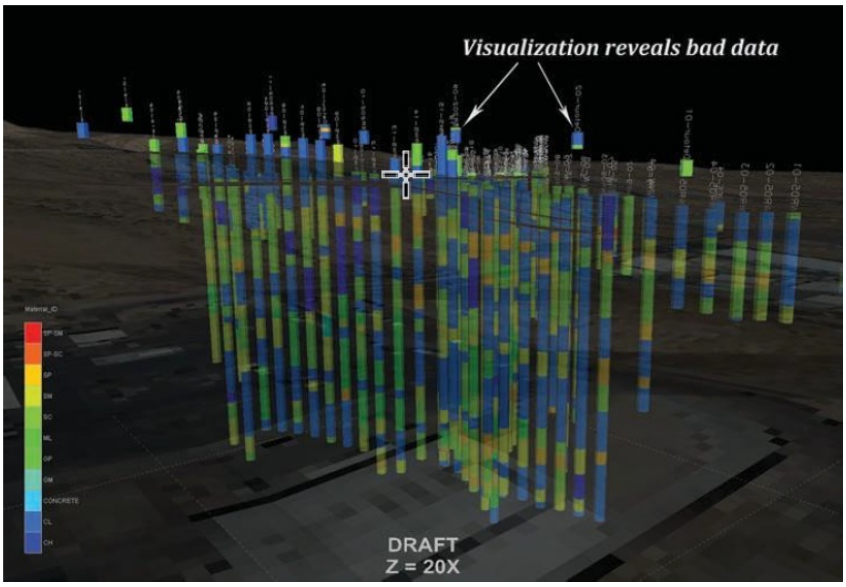


Fig. 9. Geospatial visualization is for data QA/QC. Here misplaced well data are identified by their relationship to a separate dataset – the DEM topographic dataset.

2.1 Coordinate Projections

Because the product uses EVS to create objects in a three-dimensional domain, it is necessary to have all objects projected into a consistent coordinate system. Also, because the product is designed to produce 3D CSMs for the entire state of Texas, it is preferable to use a statewide coordinate system that does not subdivide Texas into zones. For example, the popular State Plane and UTM coordinate systems are not preferable because they split the state into five and three zones, respectively. It is also not logical to use longitude and latitude values. Z values are in units of feet, and thus it is optimal to also have X and Y units in units of feet so that any volumetric calculations performed on the 3D CSMs will not require X and Y unit conversions. In addition, using longitude and latitude values wreak havoc on vertical exaggeration settings of the visualizations. Therefore, Albers Equal Area Conic was selected as the target coordinate projection for the 3D CSMs output by the product. All data will be projected as follows:

- Projection: Albers Equal Area Conic
- Units: feet
- Datum: NAD83
- Spheroid: GRS80
- 1st Std. Parallel: 27 30 00 (27.50000)
- 2nd Std. Parallel: 35 00 00 (35.00000)
- Central Meridian: -100 00 00 (-100.00000)

- Latitude of Projection: 31 15 00 (31.25000)
- False Easting: 4921250.00000 (US survey feet)
- False Northing: 19685000.00000 (US survey feet)

Each dataset obtained from the repositories come with their own dataset projections and had to be projected into the target projection. In addition, the product accepts a bounding box from a user in latitude and longitude format. The following summarizes the original coordinate projections for the datasets. The Python libraries geopandas and pyproj were used to perform the necessary projections.

2.2 Data Acquisition

The product capitalizes on three public repositories to obtain the data used to build the 3D CSMs. These three repositories are the USGS, the TWDB, and the Railroad Commission (RRC). The nature of how data is stored at these repositories varies greatly, and thus the challenge of creating a program that is able to access on-demand datasets from the sites varies greatly. In general, the RRC repositories present the greatest challenge, followed by the USGS, with the TWDB being the simplest. A summary of how the product was designed to access and download on-demand datasets from each of these repositories follows.

2.2.1 TWDB Groundwater Database

The TWDB maintains current instances of its groundwater database on their website [2]. The groundwater database contains records for approximately 140,000 water wells. For these wells, the database includes information on coordinate locations, screened intervals, water levels, and water quality, aquifer completions, well depths, surface elevations, and well types. The groundwater database is updated by the TWDB nightly and is available for download as zip file. Within the zip file are twenty pipe delimited ASCII files that can be joined based on a table attribute keys that are described in the database documentation. Figure 10 displays a snippet of part of the GWDB schema. The database format is well organized and relatively easy to import into Python for formatting and analysis and is of an easily handled size of roughly 2 GB. Because of this, the database is downloaded and unzipped daily to a local drive, and the product then accesses the data locally.

However once unpacked, and unzipped, some tables can range up to two million observations which can take some time to read in for data wrangling. Care must be taken to leverage the data in an efficient manner in order to produce the desired file formats in a reasonable amount of time. As the end user specifies a bounding box of coordinates they are interested in, paring down the data to the observations within that bounding box initially is crucial. As

[2] More information may be found at https://www.twdb.texas.gov/groundwater/data/gwdbrpt.asp. Last accessed 14 Aug 2020.

the main wells are identified within the bounding box, tables from the TWDB Groundwater data are joined by the state designated well numbers to produce a culmination of the data necessary to produce the file formats needed for the data visualization. The main attributes extracted for the EVS file formats are the state designated well number, converted longitudinal and latitudinal coordinates of the well based on the coordinate reference system, the land surface elevation of the wells, the well depths, well casing tops and bottoms, and dissolved solids, if they are present for particular wells.

This solution proves to be much simpler and faster than on-demand scraping, as must be done with the USGS data described below.

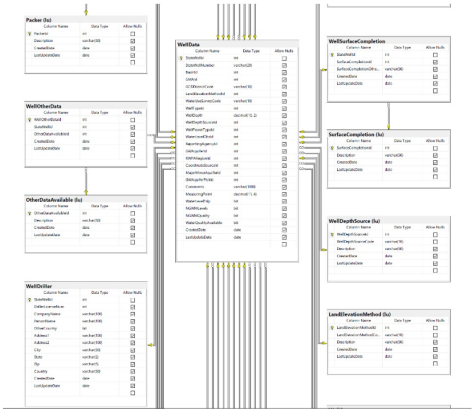


Fig. 10. Partial view of the TWDB Groundwater Database schema design.

2.2.2 USGS Data

The product obtains two datasets from the USGS repository. These are the NED DEMs and the NAIP aerial photos. Both of these datasets have national coverage, are updated regularly, and because they are high resolution, are quite large. The NAIP imagery is maintained in JPEG2000 format (Figure 11), and the DEMs are maintained as DEM raster files (Figure 12). Due to the size of these files, and the nature as to which they are updated, storing the file entire Texas datasets on a local drive or server is prohibitive. Therefore, these data are obtained by the product on-demand within the user bounding box.

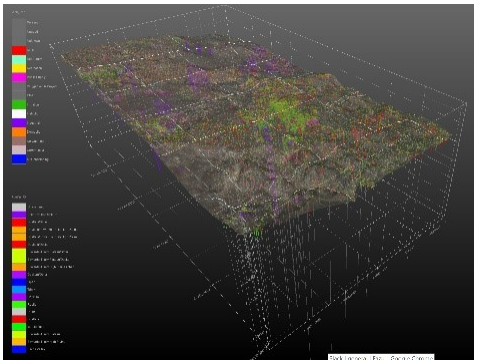


Fig. 11. View of the USGS NAIP Imagery obtained based on the parameters of the user input bounding box and visualized by the product.

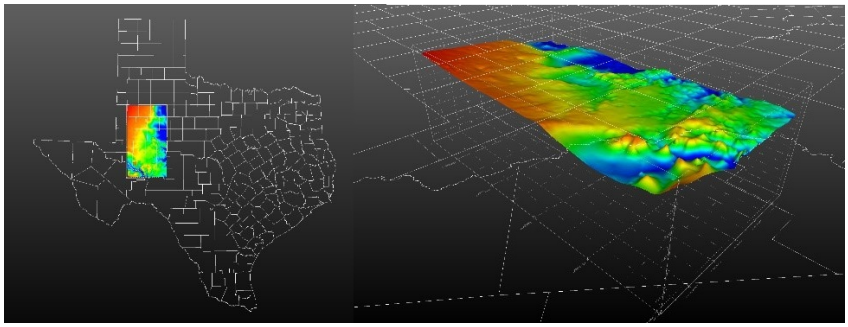


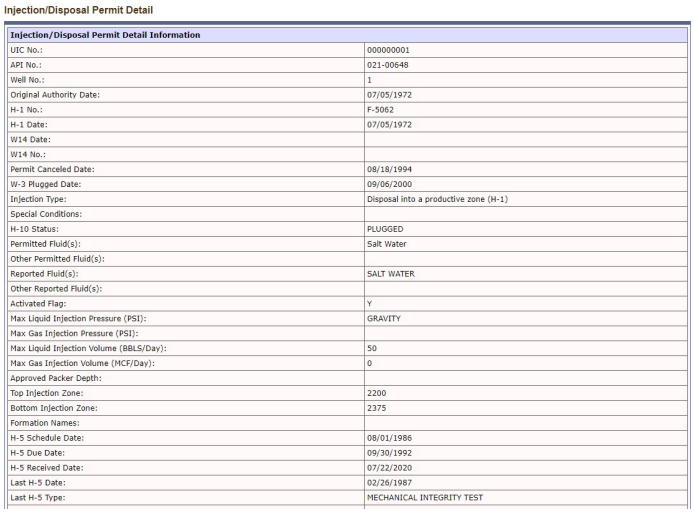
Fig. 12. View of the USGS DEM obtained based on the parameters of the user input bounding box and visualized by the product.

2.2.3 RRC Data

The data maintained by the RRC presents the greatest challenge for incorporating into an on-demand visualization system. Many of the datasets maintained by the RRC are done so on antiquated systems. First, the Underground Injection Control (UIC) represents all well data in a tabulated format built on a standard javascript-based html page. Each individual well's data is represented on its own webpage and there are currently 118,000+ injection wells. The data on these pages include:

- Well Name
- County
- District

- Depths of formation being Injected
- Type of injection
- And several more...



Injection/Disposal Permit Detail

Injection/Disposal Permit Detail Information	
UIC No.:	000000001
API No.:	021-00648
Well No.:	1
Original Authority Date:	07/05/1972
H-1 No.:	F-5062
H-1 Date:	07/05/1972
W14 Date:	
W14 No.:	
Permit Canceled Date:	08/18/1994
W-3 Plugged Date:	09/06/2000
Injection Type:	Disposal into a productive zone (H-1)
Special Conditions:	
H-10 Status:	PLUGGED
Permitted Fluid(s):	Salt Water
Other Permitted Fluid(s):	
Reported Fluid(s):	SALT WATER
Other Reported Fluid(s):	
Activated Flag:	Y
Max Liquid Injection Pressure (PSI):	GRAVITY
Max Gas Injection Pressure (PSI):	
Max Liquid Injection Volume (BBLS/Day):	50
Max Gas Injection Volume (MCF/Day):	0
Approved Packer Depth:	
Top Injection Zone:	2200
Bottom Injection Zone:	2375
Formation Names:	
H-5 Schedule Date:	08/01/1986
H-5 Due Date:	09/30/1992
H-5 Received Date:	07/22/2020
Last H-5 Date:	02/26/1987
Last H-5 Type:	MECHANICAL INTEGRITY TEST

Fig. 13. Example UIC webpage containing Injection/Disposal Permit Details.

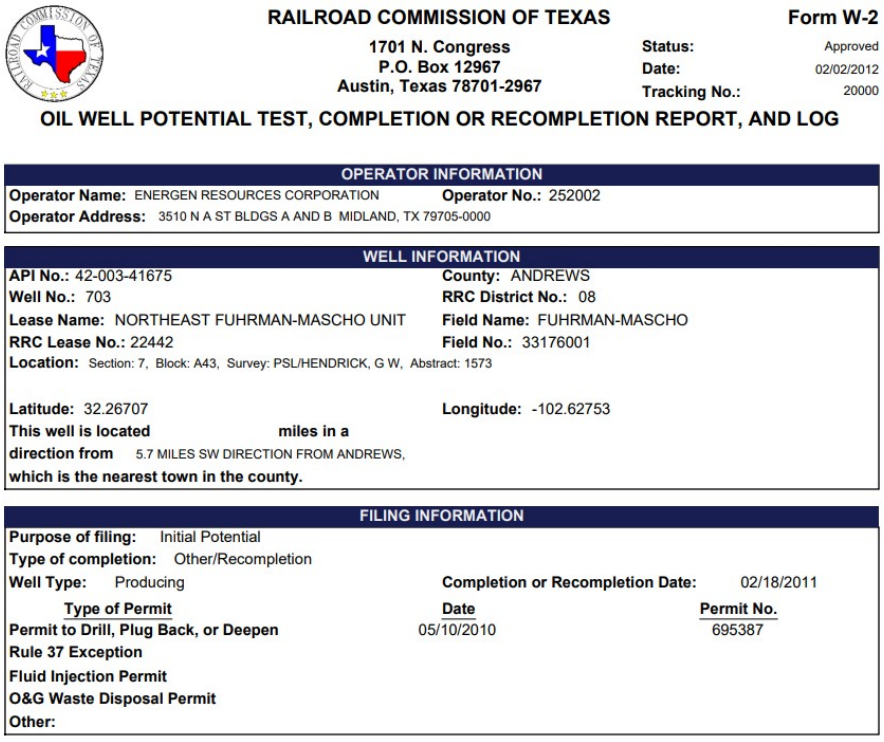
Our project utilizes asynchronous html coding and scraping techniques for this data. The asynchronous code allows to simultaneously download and table the data making the execution far much efficient. Additionally, this allows us to scrape in bulk, but also limit our volume accordingly as the RRC website does have HTTP protocols to disable mass scraping. The entire data set can be downloaded and built into a readable text file in approximately 30 hours. Updating and appending the text file as new injection wells are added to the RRC database can be done quickly adding to the existing dataset.

The UIC also has a separate database in a .dbf format which contains the lat/long data on an FTP server. The dbf files are utilized by standard database management systems. Utilizing similar scraping techniques, the dbf files were scraped into a local drive from the RRC local FTP server. This is a file server where the RRC keeps its data in various formats such as dbf and ASCII. The entire UIC data is compiled into 30 separate dbf packages that are updated monthly. We utilized dbf specific Python libraries to convert the dbf to tables and into the same format as the data scraped from the UIC. We then merged the data utilizing the API key which is a unique key provided to all Oil, Gas and Salt Water Injection wells. We can then export to a .txt file which is ready to be used.

Finally, the oil and gas wells are represented on a W-2 report. The W-2 is a state regulated form all oil and gas companies fill out and submit to the state for

each well that is "completed" and producing hydrocarbons. This is a designation for wells that have been hydraulically fractured which is a technique employed by oil and gas companies. The W-2 is presented in an embedded pdf within the javascript html. There are currently over 250,000 oil and gas wells with W-2 pdfs that can be obtained. The data obtained from these are incredibly useful for visualization which include:

- Producing Formations (where hydrocarbons are extracted from)
- Formation tops for all geologically defined depths
- Lat/longs
- Detailed well information such as depth, mechanical components and more. . .



RAILROAD COMMISSION OF TEXAS — **Form W-2**

1701 N. Congress
P.O. Box 12967
Austin, Texas 78701-2967

Status: Approved
Date: 02/02/2012
Tracking No.: 20000

OIL WELL POTENTIAL TEST, COMPLETION OR RECOMPLETION REPORT, AND LOG

OPERATOR INFORMATION

Operator Name: ENERGEN RESOURCES CORPORATION **Operator No.:** 252002
Operator Address: 3510 N A ST BLDGS A AND B MIDLAND, TX 79705-0000

WELL INFORMATION

API No.: 42-003-41675 **County:** ANDREWS
Well No.: 703 **RRC District No.:** 08
Lease Name: NORTHEAST FUHRMAN-MASCHO UNIT **Field Name:** FUHRMAN-MASCHO
RRC Lease No.: 22442 **Field No.:** 33176001
Location: Section: 7, Block: A43, Survey: PSL/HENDRICK, G W, Abstract: 1573

Latitude: 32.26707 **Longitude:** -102.62753
This well is located **miles in a**
direction from 5.7 MILES SW DIRECTION FROM ANDREWS,
which is the nearest town in the county.

FILING INFORMATION

Purpose of filing: Initial Potential
Type of completion: Other/Recompletion
Well Type: Producing **Completion or Recompletion Date:** 02/18/2011

Type of Permit	Date	Permit No.
Permit to Drill, Plug Back, or Deepen	05/10/2010	695387
Rule 37 Exception		
Fluid Injection Permit		
O&G Waste Disposal Permit		
Other:		

Fig. 14. Example W-2 report.

By using a pinpointed web scraping technique, the embedded pdf can be downloaded to a local drive/server, or database format of the user's choosing. The base code will download the pdf's in bulk, scrape the pdf and table the data into a useable format. It will also delete those pdfs (if desired). The code can

also append the dataset as new oil and gas wells are added to the RRC database. Downloading the initial volume of wells took approximately 5 days. Updating on a go-forward basis takes minutes. There are several other pdfs, dbfs and other data sources from the RRC that can be utilized through these scraping and munging techniques. The potential to grow our database to show users both volume and quality of data down to a very granular level is vast. Figure 15 shows an example of the total depth of vertical wells, and the total depth and transects of horizontal wells obtained from within the parameters of a user specified bounding box and visualized by the product. Figure 16 shows RRC UIC well data, along with injection interval and injection volume attributes along with TWDB aquifer layers. This sort of geospatial data combinations are particularly valuable for investigating relationships between the otherwise disparate data sources (i.e. Texas groundwater data and Texas oil and gas data).

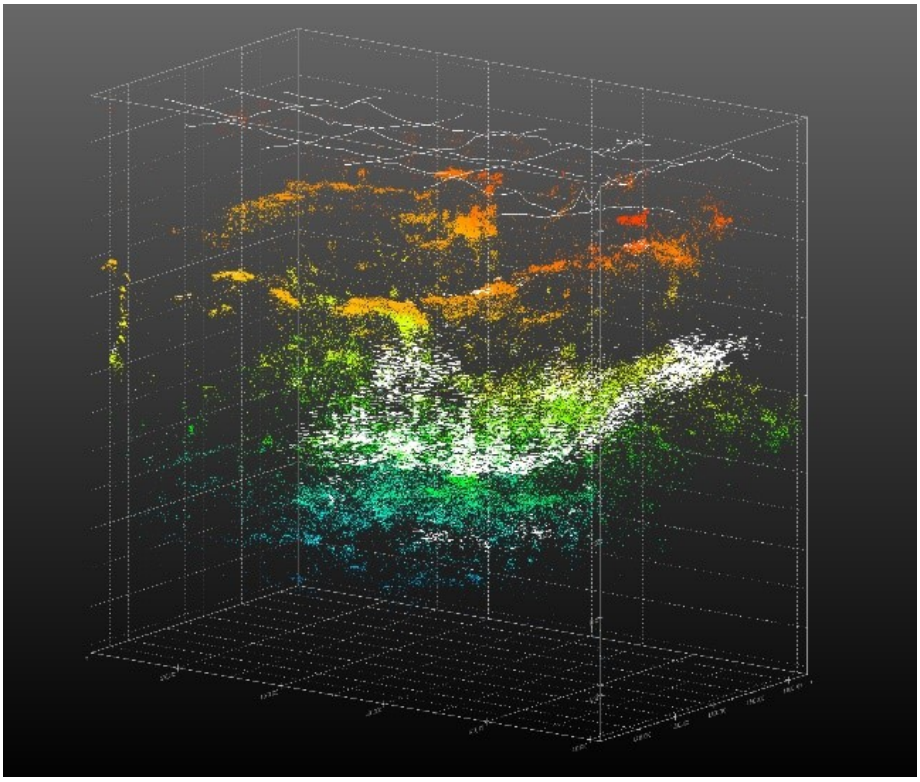


Fig. 15. View of Texas Railroad Commission oil and gas well data and attributes obtained based on the parameters of the user input bounding box and visualized by the product.

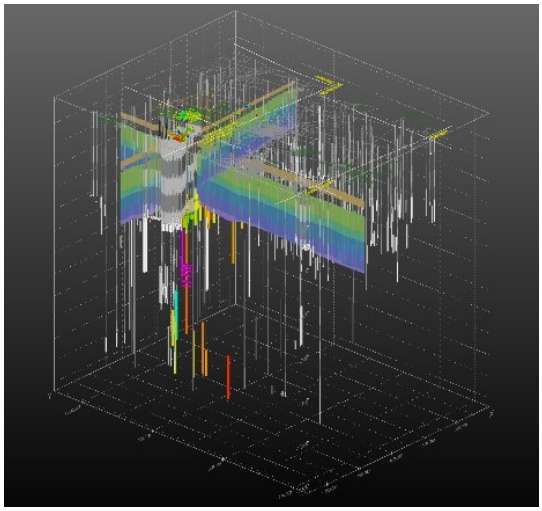


Fig. 16. View of Texas Railroad Commission underground injection control (UIC) data and attributes obtained based on the parameters of the user input bounding box and visualized by the product. Also shown are slices of TWDB obtained aquifer layer surfaces. This sort of geospatial data combinations are particularly valuable for investigating relationships between the otherwise disparate data sources (i.e. Texas groundwater data and Texas oil and gas data).

2.3 Data Formatting

EVS provides detailed documentation on acceptable file formats in its documentation, included with the software and available online at their website [3].

In order for the product to use EVS for data visualization, the data first has to be formatted into the appropriate EVS file format. Depending on which dataset one is dealing with, this can be a simple, or quite complicated task. The product utilizes the following EVS file formats: Borehole Geology Format (.geo), 3DAnalyte Format (.apv), and 3D Groundwater Analyte Format (.aidv), Geology Mult-File Format (.gmf), 2D and 3D shape files and rasters.

3 Ethics

As the data that is collected for this product is available publicly, the concern for data security was not considered. Personal data is not utilized in any of the data visualization models, so extra consideration of securing the data will not be performed at this time. If any new data sources are added or existing data sources move from a public domain to a private domain, the security of the data will need to be evaluated and handled accordingly.

[3] More information may be found at https://www.ctech.com. Last accessed 14 Aug 2020.

As errors or oversights in the data are found, it is our ethical obligation to inform the organizations who provide the data in order to ensure that anyone accessing the data has the most accurate data possible at hand. Identifying inaccurate data not only helps our product and end users but also benefits the entire industry to help preserve the environment and minimizes waste or excessive work needed to be performed. Testing and exploratory analysis will also be performed on the data to verify the quality of the data acquired and to identify any outliers that may be possible errors.

Although the goal of the product is to provide end users with as much information about a site through data visualization, care must be taken to ensure the visualizations are not misleading. Deceptive data visualization is a general issue today that is prevalent and being monitored more and more. Any misleading data visualizations would be unintentional as we are striving to provide as much data about a site with the objective of high accuracy, but we will periodically review generated visualizations for any discrepancies or areas that may be misleading. Additionally, there will be gaps in data which may require imputation. Any methods of imputation will be made clear to end users to provide clarity. In the case of geospatial visualizations, imputations of data can be made in several ways, such as average the distance and slope of formations as the geologically dip from well to well. Alternatively, estimations of dips in formations can be done by estimating pinch out or expansion of the actual formation thickness. These are important distinctions as geologists and engineers may view the methodology differently for each imputation method.

Being that we are creating a product for commercial consumption, care must be taken to regularly test the code throughout the development process to ensure we are providing an accurate and high-quality product. Creating a suite of tests that cover various aspects of the code base such as data acquisition, data processing, file generation are some example areas that should be implemented when possible. Once the series of tests are in place, the next step would be to automate the testing process and run it regularly as a means to regression test the product to help minimize the introduction of software bugs released into production. If any software bug issues are reported or identified by end users, the issues will be investigated and confirmed if the issue does indeed exist. Once the issue is confirmed, the steps to remedy the bug will be taken and a new version of the software will be released. To help ensure that the issue does not occur again, a test for that issue will be added to the automated regression test suite. As more and more issues are identified and repaired, the growing automated test suite will further improve upon the quality of the product and help prevent introducing the same issues in the future.

4 Summary and Conclusions

Subsurface models are an essential component of Geoscience and Geoscience consultation. Our automated process reduces the considerable amount of latency that was previously unavoidable. By automating the data acquisition, reformat-

ting the data, and implementation into EVS, what took weeks to build can now be done in minutes. Our project is in the prototype phase and is already useable for business purposes. Although not fully finished, it has already been implemented in assisting clients. It resolves the issues with inconsistent and bad data, prevents revenue loss and misused resources. With having the data centralized, further analysis can be performed beyond what even this project entails. Acquiring this data through third party services is costly and is typically charged as a monthly subscription. As new wells are looked upon by Exploration companies, they can utilize the well data provided to analyze offset wells to inform them to make better decisions. This includes drilling depth, collision avoidance and revenue loss prevention by avoiding water injection wells. Exploration companies can also maintain regulatory compliance with the Texas Water Board by using this to determine accurate aquifer depths in their assets. In conclusion, Automated Geospatial Data Visualization has a multitude of use cases in the real world and creates immediate value for end users.